**Northeastern University**

**College of Professional Studies**



**ALY-6040 Data Mining Applications**

**TOPIC:** **Finding Patterns in Data and EDA**

**TO: Prof. Justin Grosz**

**BY:** Snehal Keshav Bende

Khushboo Bharatkumar Oza

Shruti Sawale

Mohit Patel

Rishi Barbhaya

**Quarter 2- WINTER 2020**

**Bank Marketing campaign**

Bank marketing is an important factor to get potential customers by focusing on their needs and satisfaction. The campaign involves various strategies to attract more clients such as population which will tell which category of population is more involved, price as the offer should be worth the money, where the campaign is taking place and how, implementing strategies that will rectify previous mistakes and improve customer relationship. The strategies could involve a lot of parameters like the education and occupation of the customer, loans that the customer has, the month that he is being contacted in, the number of times a customer has to be called for pursuing him for term deposit and follow ups.

We wanted to learn how bank formulates its strategies to offer customers customized products according to different factors like loans, age, occupation and so on which lead us to the selection of this dataset.

The objective of our analysis is to increase the term depositors by enhancing the marketing strategies. Eventually, this will also result in success of the bank.

**DATA:**

Source: Martinez, J. (2017, November 12). Bank Marketing Dataset. Retrieved from <https://www.kaggle.com/janiobachmann/bank-marketing-dataset>.

The dataset consists of 17 attributes and 11162 rows. The target variable is ‘deposit’ which is categorial attribute with values ‘yes’ and ‘no’.

The remaining attribute are as follows:

Age: numeric data with range 18-95

Job: categorical, which has types such as admin, blue collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, un-employed, unknown

Marital: categorical, status: divorced (note: 'divorced' means divorced or widowed), married, single

Education: categorical, type: primary, secondary, tertiary and unknown

Default: categorical, yes or no. Determines whether person has credit in default or not

Housing: categorical, value: yes, no. Determines whether person has housing loan or not.

Loan: categorical, value: yes, no. Determines whether person has personal loan or not.

Balance: numeric, individuals balance.

Contact: categorical, type of communication: cellular and telephone

Month: categorical, last month when person was contacted

Day: numeric, last day when the person was contacted. Range 1- 31

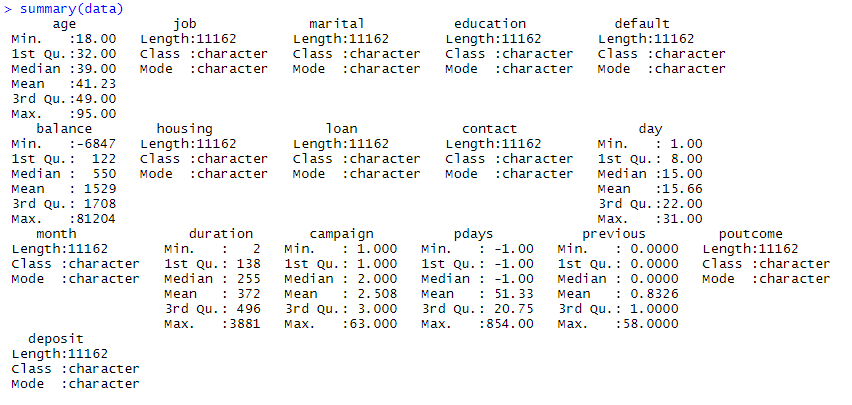
Duration: numeric, the last contact duration.

Campaign: numeric, number of contacts performed for client for this campaign

Pdays: numeric, number of days passed after contacting the client

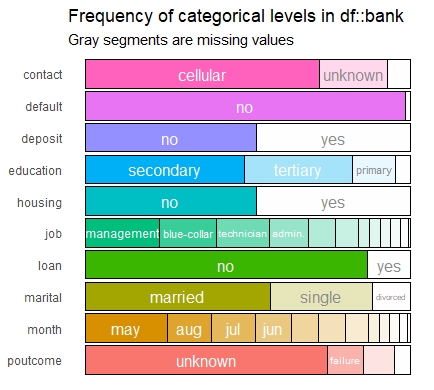
Previous: numeric, number of contacts performed for previous campaign

Poutcome: categorical, outcome of the previous marketing campaign ( 'failure','nonexistent','success')



**DATA CLEANING**

The first step of the process was to import the data which was in csv file. After importing the file, in the dataset the first step was to find out NULL and MISSING values. The dataset didn’t have any of these values but had many ‘unknown’ values. The attribute ‘poutcome’ consisted of 74% data of ‘unknown’ values hence the column was deleted. It is observed that the target variable is balanced which means that no resampling was required. Also, column ‘education’ consisted 4.4% and column ‘job’ consisted of 0.6% data of ‘unknown’ values, these rows were deleted as the percentage of ‘unknown’ value is comparatively less. Also, for future analysis we have converted some of the categorical values to numeric values. Attribute ‘contact’ also had 21% of ‘unknown’ values as the percentage was moderate and as well as there can be other modes of contacting like emails, personal meets, etc., these values cannot be ignored here. The same can be inferred from the below graph.



**EDA:**

For data exploration, relationship between the attributes were found and how they can affect the marketing strategies of campaigning was observed. The correlation graph below depicts correlation of various attributes with each other and it was observed that few of them have negative correlation. Negative correlation signifies that increase in one attribute leads to decrease in other attribute which in a way means they are inversely proportional.

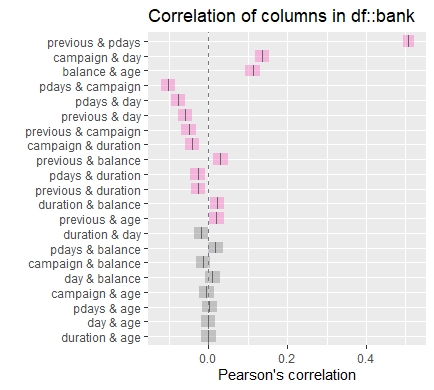
The following are negative correlations observed:

* pdays and campaign
* Pdays and day
* Previous and day
* Previous and campaign
* Campaign and duration
* Pdays and duration
* Previous and duration

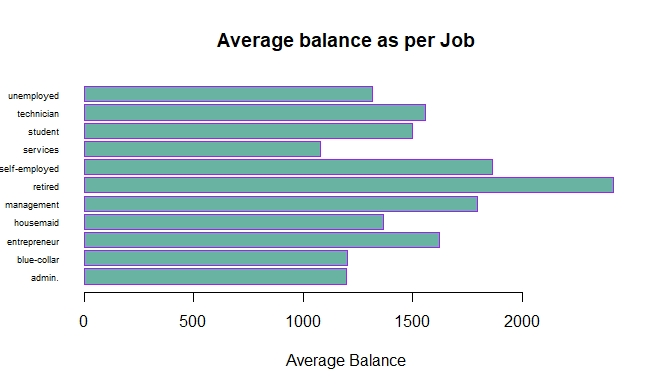
The following are positive correlations observed:

* Previous and pdays
* Campaign and days
* Balance and age
* Previous and balance
* Duration and balance
* Previous and age

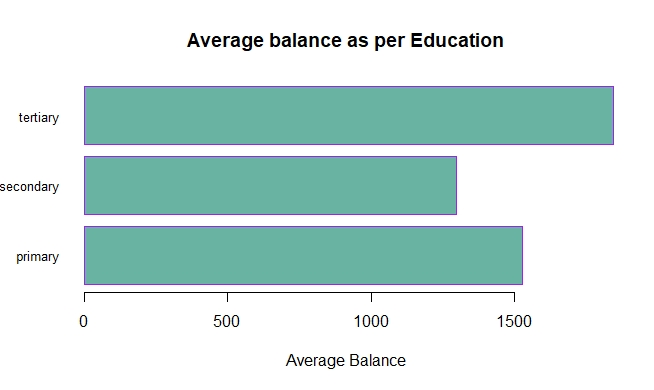
The maximum negative correlation was observed between pdays and campaign, whereas maximum positive correlation was observed between previous and pdays.



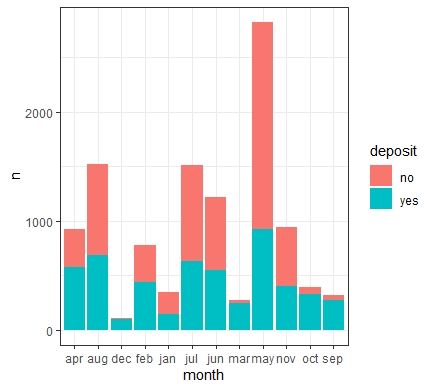
Average balance was calculated by grouping the jobs of client and it was observed that retired category holds the maximum balance in the bank and services holds the minimum. This in turn tells which category should be targeted for campaign.



The below graph shows that average balance of education for each category. People with tertiary education hold maximum balance which could be considered while giving the loans or can make the repayment.



A bar graph of months vs number of deposits was plotted to check which months have highest depositors. It was noted that ‘MAY’ has highest, and December has lowest.



The findings obtained from the above EDA helped us analyze that which month had maximum depositors and identify campaign needs to be done to increase the number of depositors in rest of the months as well. From the ‘average balance as per education’ graph it can also be noted that tertiary education plays an important role as people with that level of education have maximum balance which in turn will also lead to potential term depositors .These finding play a major part of forming a strategy for campaigning as it tells which attributes and their relationship to focus on for maximum effectiveness.

The next steps would be finding out potential candidates that will belong to certain occupation, education and age category. Also figure out what should be the policies for campaign calls. Lastly building that will help to predict potential term depositors.

